

Randomized Block Kaczmarz Method with Projection for Solving Least Squares

Deanna Needell^{a,*}, Ran Zhao^b, Anastasios Zouzias^c

^aDept. of Mathematical Sciences, Claremont McKenna College, Claremont, CA 91711

^bDept. of Mathematics, Claremont Graduate Univ., Claremont, CA 91711

^cIBM Research Lab, Zurich

Abstract

The Kaczmarz method is an iterative method for solving overcomplete linear systems of equations $\mathbf{Ax} = \mathbf{b}$. The randomized version of the Kaczmarz method put forth by Strohmer and Vershynin iteratively projects onto a randomly chosen solution space given by a single row of the matrix \mathbf{A} and converges linearly¹ in expectation to the solution of a *consistent* system. In this paper we analyze two block versions of the method each with a randomized projection, designed to converge in expectation to the least squares solution, often faster than the standard variants. Our approach utilizes both a row and column-paving of the matrix \mathbf{A} to guarantee linear convergence when the matrix has consistent row norms (called nearly standardized), and a single column-paving when the row norms are unbounded. The proposed methods are an extension of the block Kaczmarz method analyzed by Needell and Tropp and the Randomized Extended Kaczmarz method of Zouzias and Freris. The contribution is thus two-fold; unlike the standard Kaczmarz method, our results demonstrate convergence to the least squares solution of inconsistent systems (both methods in the nearly standardized case and the second method in other cases). By using appropriate blocks of the matrix this convergence can be significantly accelerated, as is demonstrated by numerical experiments.

Keywords: block Kaczmarz, randomized extended Kaczmarz, projections onto convex sets, algebraic reconstruction technique, matrix paving

2000 MSC: 65F10, 65F20, 68W20, 41A65

1. Introduction

The Kaczmarz method [Kac37] is a popular iterative solver of overdetermined systems of linear equations $\mathbf{Ax} = \mathbf{b}$. Because of its simplicity and performance, the method and its

*Corresponding Author

Email addresses: dneedell@cmc.edu (Deanna Needell), ran.zhao@cgu.edu (Ran Zhao), azo@zurich.ibm.com (Anastasios Zouzias)

¹Mathematicians often refer to this type of convergence as exponential.

derivatives are used in a range of applications from image reconstruction to digital signal processing [CFM⁺92, Nat01, SS87]. The method performs a series of orthogonal projections and iteratively converges to the solution of the system of equations. It is therefore computationally feasible even for very large and overdetermined systems.

Given a vector \mathbf{b} and an $n \times d$ full rank (real or complex) matrix \mathbf{A} with rows $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$, the algorithm begins with an initial estimate \mathbf{x}_0 and cyclically projects the estimation onto each of the solution spaces. This process can be described as follows:

$$\mathbf{x}_j = \mathbf{x}_{j-1} + \frac{\mathbf{b}[i] - \langle \mathbf{a}_i, \mathbf{x}_{j-1} \rangle}{\|\mathbf{a}_i\|_2^2} \mathbf{a}_i,$$

where $\mathbf{b}[i]$ denotes the i th coordinate of \mathbf{b} , \mathbf{x}_j is the estimation in the j th iteration, $\|\cdot\|_2$ denotes the usual ℓ_2 vector norm, $\langle \cdot, \cdot \rangle$ the standard ℓ_2 inner product, and $i = (j \bmod n) + 1$ cycles through the rows of \mathbf{A} .

Since the method cycles through the rows of \mathbf{A} , the performance of the algorithm may depend heavily on the ordering of these rows. A poor ordering may lead to very slow convergence. To overcome this obstacle, one can select the row \mathbf{a}_i at random to improve the convergence rate [HM93, Nat01]. Strohmer and Vershynin proposed and analyzed a method which selects a given row with probability proportional to its ℓ_2 norm [SV09, SV06]. They show that with this selection strategy, the randomized Kaczmarz method has an expected linear convergence rate to the unique solution² \mathbf{x}_\star :

$$\mathbb{E} \|\mathbf{x}_j - \mathbf{x}_\star\|_2^2 \leq \left(1 - \frac{1}{R}\right)^j \|\mathbf{x}_0 - \mathbf{x}_\star\|_2^2, \quad (1)$$

where R is the scaled condition number, $R = \|\mathbf{A}^{-1}\|^2 \|\mathbf{A}\|_F^2$, $\|\cdot\|_F$ denotes the Frobenius norm, and $\|\mathbf{A}^{-1}\| \stackrel{\text{def}}{=} \inf\{M : M\|\mathbf{Ax}\|_2 \geq \|\mathbf{x}\|_2 \text{ for all } \mathbf{x}\}$ is well-defined since \mathbf{A} has full column rank. This convergence rate (1) is essentially independent of the number of rows of \mathbf{A} and shows that for well-conditioned matrices, the randomized Kaczmarz method converges to the solution in just $O(d)$ iterations [SV09]. This yields an overall runtime of $O(d^2)$ which is much superior to others such as $O(nd^2)$ for Gaussian elimination. There are also cases where randomized Kaczmarz even outperforms the conjugate gradient method, see the discussion in [SV09] for details.

When the system is perturbed by noise or no longer consistent, $\mathbf{Ax}_\star + \mathbf{e} = \mathbf{b}$, the randomized Kaczmarz method still provides expected linear convergence down to an error threshold [Nee10],

$$\mathbb{E} \|\mathbf{x}_j - \mathbf{x}_\star\|_2 \leq \left(1 - \frac{1}{R}\right)^{j/2} \|\mathbf{x}_0 - \mathbf{x}_\star\|_2 + \sqrt{R} \cdot \max_i \frac{|e[i]|}{\|\mathbf{a}_i\|_2}, \quad (2)$$

²Note that if the full-rank assumption is removed, the method converges to the solution set at the same linear rate with $\|\mathbf{A}^{-1}\|$ replaced by the reciprocal of the smallest non-zero singular value. See e.g. (3) of [LWS14].

where $|e[i]|$ denotes the i th entry of \mathbf{e} . This result is sharp, and shows that the randomized Kaczmarz method converges with a radius proportional to the magnitude of the largest entry of the noise in the system. Since the iterates of the Kaczmarz method always lie in a single solution space, the method clearly will not converge to the least squares solution of an inconsistent system.

1.1. Randomized Extended Kaczmarz

The bound (2) demonstrates that the randomized Kaczmarz method performs well when the noise in inconsistent systems is small. Zouzias and Freris introduced a variant of the method which utilizes a random projection to iteratively reduce the norm of the error [ZF12]. They show that the estimate of this *Randomized Extended Kaczmarz* (REK) method converges linearly in expectation to the least squares solution of the system, breaking the radius barrier of the standard method. The algorithm maintains not only an estimate \mathbf{x}_j to the solution but also an approximation \mathbf{z}_j to the projection of \mathbf{b} onto the range of \mathbf{A} :

$$\mathbf{x}_j = \mathbf{x}_{j-1} + \frac{\mathbf{b}[i] - \mathbf{z}_{j-1}[i] - \langle \mathbf{a}_i, \mathbf{x}_{j-1} \rangle}{\|\mathbf{a}_i\|_2^2} \mathbf{a}_i, \quad \mathbf{z}_j = \mathbf{z}_{j-1} - \frac{\langle \mathbf{a}^{(k)}, \mathbf{z}_{j-1} \rangle}{\|\mathbf{a}^{(k)}\|_2^2} \mathbf{a}^{(k)}, \quad (3)$$

where in iteration j , \mathbf{a}_i and $\mathbf{a}^{(k)}$ is the row and column of \mathbf{A} , respectively, each chosen randomly with probability proportional to their Euclidean norms. In this setting, we no longer require that the matrix \mathbf{A} be full rank, and ask for the least squares solution,

$$\mathbf{x}_{LS} \stackrel{\text{def}}{=} \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 = \mathbf{A}^\dagger \mathbf{b},$$

where \mathbf{A}^\dagger denotes the Moore-Penrose pseudoinverse of \mathbf{A} . Zouzias and Freris showed that the REK method converges linearly in expectation to the least squares solution [ZF12],

$$\mathbb{E} \|\mathbf{x}_j - \mathbf{x}_{LS}\|_2^2 \leq \left(1 - \frac{1}{K^2(\mathbf{A})}\right)^{j/2} \left(\|\mathbf{x}_{LS}\|_2^2 + \frac{2\|\mathbf{b}\|_2^2}{\sigma_{\min}^2(\mathbf{A})} \right), \quad (4)$$

where $\sigma_{\min}(\mathbf{A})$ is the smallest non-zero singular value of \mathbf{A} and $K(\mathbf{A}) = \frac{\|\mathbf{A}\|_F}{\sigma_{\min}(\mathbf{A})}$ denotes its scaled condition number.

1.2. The block Kaczmarz method

Recently, Needell and Tropp analyzed a block version of the simple randomized Kaczmarz method [NT13]. Like the traditional method, this version iteratively projects the current estimation onto the solution spaces. However, rather than using the solution space of a *single* equation, the block method projects onto the solution space of many equations simultaneously by selecting a block of rows rather than a single row. For a subset $\tau \subset \{1, 2, \dots, n\}$, denote by \mathbf{A}_τ the submatrix of \mathbf{A} whose rows are indexed by τ . We again begin with an arbitrary guess \mathbf{x}_0 for the solution of the system. Then for each iteration $j \geq 1$,

select a block $\tau = \tau_j$ of rows. To obtain the next iterate, we project the current estimation onto the solution space of the equations listed in τ [NT13]:

$$\mathbf{x}_j = \mathbf{x}_{j-1} + (\mathbf{A}_\tau)^\dagger (\mathbf{b}_\tau - \mathbf{A}_\tau \mathbf{x}_{j-1}). \quad (5)$$

Here, the conditioning of the blocks \mathbf{A}_τ plays a crucial role in the behavior of the method. Indeed, if each block is well-conditioned, its pseudoinverse can be applied efficiently using an iterative method such as CGLS [Bjö96]. To guarantee such properties, Needell and Tropp utilize a *paving* of the matrix \mathbf{A} .

Definition 1 (Row Paving). A (p, α, β) row paving of a $n \times d$ matrix \mathbf{A} is a partition $\mathcal{T} = \{\tau_1, \dots, \tau_p\}$ of the rows such that

$$\alpha \leq \lambda_{\min}(\mathbf{A}_\tau \mathbf{A}_\tau^*) \quad \text{and} \quad \lambda_{\max}(\mathbf{A}_\tau \mathbf{A}_\tau^*) \leq \beta \quad \text{for each } \tau \in \mathcal{T},$$

where again we denote by \mathbf{A}_τ the $|\tau| \times d$ submatrix of \mathbf{A} . We refer to the number p as the size of the paving, and the numbers α and β are called the lower and upper paving bounds.

We refer to a row paving of \mathbf{A}^* as a *column paving* of \mathbf{A} . We thus seek pavings of \mathbf{A} with small number of blocks p and upper paving constant β . In the following, we will assume one has access to such a paving, and discuss in Section 3 how to construct such pavings for various types of matrices. When the matrix has unit-norm rows, equipped with such a row paving of \mathbf{A} , the main result of [NT13] shows that the randomized block Kaczmarz algorithm (5) exhibits linear convergence in expectation:

$$\mathbb{E} \|\mathbf{x}_j - \mathbf{x}_\star\|_2^2 \leq \left(1 - \frac{1}{C' \kappa^2(\mathbf{A}) \log(1+n)}\right)^j \|\mathbf{x}_0 - \mathbf{x}_\star\|_2^2 + \frac{3\|\mathbf{e}\|_2^2}{\sigma_{\min}^2(\mathbf{A})}, \quad (6)$$

where C' is an absolute constant and $\kappa(\mathbf{A}) = \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$ is the condition number of \mathbf{A} .

Since each iteration of the block method utilizes multiple rows, one can compare the rate of (6) and (1) by considering convergence per epoch (one cycle through the rows of \mathbf{A}). From this analysis, one finds the bounds to be comparable. However, the block method can utilize fast matrix multiplies and efficient implementation, yielding dramatic improvements in computational time. See [NT13] for details and empirical results.

1.3. Contribution

The REK method breaks the so-called *convergence horizon* of standard Kaczmarz method, allowing convergence to the least squares solution of inconsistent systems. The block Kaczmarz method on the other hand, allows for significant computational speedup and accelerated convergence to within a fixed radius of the least squares solution. The main contribution of this paper analyzes a randomized block Kaczmarz method which also incorporates a blocked projection step, which provides accelerated convergence to the least squares solution. In this case we need a column partition for the projection step and a row partition for the Kaczmarz step. Our results show that this method offers both linear

convergence to the least squares solution \mathbf{x}_{LS} when a row paving can be obtained, and improved convergence speed due to the blocking of both the rows and columns. In addition, we present a block coordinate descent variant which utilizes only a column paving and also yields linear convergence. We will see that the desired column paving may be obtained easily for arbitrary matrices, so this variant is especially useful when a good row paving cannot be obtained.

1.4. Organization

In Section 2 we begin by presenting the double block randomized Kaczmarz method which utilizes both a row and column paving, and provides the strongest theoretical results overall. We discuss methods for obtaining the desired matrix pavings in Section 3. Section 4 introduces the variant of the block and extended Kaczmarz methods that requires only a column paving of the matrix, readily accessible for arbitrary matrices. In Section 6 we present some experimental results for the various algorithms. We conclude with a discussion of related work and open directions in Section 7. The appendix includes proofs of intermediate results used along the way.

2. The Randomized Double Block Kaczmarz Method

It is natural to ask whether one can consider both a row partition and a column partition in the Kaczmarz method, blocking both in the Kaczmarz update step and the projection step. Indeed, utilizing blocking in both steps yields Algorithm 1 below. We thus propose the following randomized block extended Kaczmarz method using double partitioning. We will see later that since this method requires a row paving, it works very well when each of the row norms of the matrix are relatively similar (see Sections 3 and 5).

Algorithm 1 Randomized Double Block Kaczmarz Least Squares Solver

```

1: procedure ( $A, \mathbf{b}, T, \mathcal{T}, \mathcal{S}$ )  $\triangleright A \in \mathbb{R}^{n \times d}, \mathbf{b} \in \mathbb{R}^n, T \in \mathbb{N}$ , column partition  $\mathcal{T}$  of  $[d]$ , row
   partition  $\mathcal{S}$  of  $[n]$ 
2:   Initialize  $\mathbf{x}_0 = \mathbf{0}$  and  $\mathbf{z}_0 = \mathbf{b}$ 
3:   for  $k = 1, 2, \dots, T$  do
4:     Pick  $\tau_k \in \mathcal{T}$  and  $v_k \in \mathcal{S}$  uniformly at random
5:     Set  $\mathbf{z}_k = \mathbf{z}_{k-1} - A_{\tau_k} (A_{\tau_k})^\dagger \mathbf{z}_{k-1}$   $\triangleright A_{\tau_k}: n \times |\tau_k|$  submatrix of  $A$ 
6:     Update  $\mathbf{x}_k = \mathbf{x}_{k-1} + (A_{v_k})^\dagger (\mathbf{b}_{v_k} - (\mathbf{z}_k)_{v_k} - A_{v_k} \mathbf{x}_{k-1})$   $\triangleright A_{v_k}: |v_k| \times d$  submatrix of  $A$ 
7:   end for
8:   Output  $\mathbf{x}_T$ 
9: end procedure

```

Combining the theoretical approaches in [NT13, ZF12] we will prove the following result about the convergence of Algorithm 1. This result utilizes both a column paving and row paving.

Theorem 1. Algorithm 1 with input \mathbf{A} , \mathbf{b} , $T \in \mathbb{N}$, $(\bar{p}, \bar{\alpha}, \bar{\beta})$ column paving \mathcal{T} of \mathbf{A} , and (p, α, β) row paving \mathcal{S} of \mathbf{A} , outputs an estimate vector \mathbf{x}_T that satisfies

$$\mathbb{E} \|\mathbf{x}_T - \mathbf{x}_{LS}\|_2^2 \leq \gamma^T \|\mathbf{x}_0 - \mathbf{x}_{LS}\|_2^2 + \left(\gamma^{\lfloor T/2 \rfloor} + \bar{\gamma}^{\lfloor T/2 \rfloor} \right) \frac{\|\mathbf{b}_{\mathcal{R}(\mathbf{A})}\|_2^2}{\alpha(1-\gamma)},$$

where $\gamma = 1 - \frac{\sigma_{\min}^2(\mathbf{A})}{p\beta}$ and $\bar{\gamma} = 1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\bar{p}\bar{\beta}}$.

Proof. We begin with the following lemma which is motivated by Lemma 2.2 of [NT13], and shows that the iterates \mathbf{z}_k converge linearly to the projection of \mathbf{b} onto the kernel of \mathbf{A}^* .

Lemma 2. Let \mathbf{b} be a fixed vector and \mathcal{T} be a $(\bar{p}, \bar{\alpha}, \bar{\beta})$ column paving of \mathbf{A} . Assuming the notation of Theorem 1, for every $k > 0$ it holds that

$$\mathbb{E} \|\mathbf{z}_k - \mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\bar{p}\bar{\beta}} \right)^k \|\mathbf{b}_{\mathcal{R}(\mathbf{A})}\|_2^2. \quad (7)$$

where $\mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp} := (\mathbf{I} - \mathbf{A}\mathbf{A}^\dagger)\mathbf{b}$.

Proof. Let $\mathbf{P}_{\tau_k} = \mathbf{A}_{\tau_k}(\mathbf{A}_{\tau_k})^\dagger$ and notice $\mathbf{z}_k = (\mathbf{I} - \mathbf{P}_{\tau_k})\mathbf{z}_{k-1}$. Define $\mathbf{e}_k = \mathbf{z}_k - \mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp}$ for $k \geq 0$. Then,

$$\begin{aligned} \mathbf{e}_k &= (\mathbf{I} - \mathbf{P}_{\tau_k})\mathbf{z}_{k-1} - \mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp} \\ &= (\mathbf{I} - \mathbf{P}_{\tau_k})\mathbf{z}_{k-1} - (\mathbf{I} - \mathbf{P}_{\tau_k})\mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp} \\ &= (\mathbf{I} - \mathbf{P}_{\tau_k})\mathbf{e}_{k-1}, \end{aligned}$$

where the first equality follows by the definition of \mathbf{z}_k , the second by orthogonality between the range of \mathbf{P}_{τ_k} and $\mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp}$, and the final equality by definition of \mathbf{e}_{k-1} . Next, we prove that

$$\mathbb{E}_{k-1} \|\mathbf{e}_k\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\bar{p}\bar{\beta}} \right) \|\mathbf{e}_{k-1}\|_2^2$$

where \mathbb{E}_{k-1} is the expectation conditioned over the first $(k-1)$ iterations of the algorithm. By orthogonality of the projector \mathbf{P}_{τ_k} and the Pythagorean theorem, one has $\|(\mathbf{I} - \mathbf{P}_{\tau_k})\mathbf{e}_{k-1}\|_2^2 = \|\mathbf{e}_{k-1}\|_2^2 - \|\mathbf{P}_{\tau_k}\mathbf{e}_{k-1}\|_2^2$, hence it suffices to lower bound $\mathbb{E}_{k-1} \|\mathbf{P}_{\tau_k}\mathbf{e}_{k-1}\|_2^2$. Let $\mathbf{A}_{\tau_k} := \mathbf{U}_{\tau_k} \mathbf{\Sigma}_{\tau_k} \mathbf{V}_{\tau_k}^*$ be the truncated SVD decomposition of \mathbf{A}_{τ_k} where $\mathbf{\Sigma}_{\tau_k}$ is a $\text{rank}(\mathbf{A}_{\tau_k}) \times \text{rank}(\mathbf{A}_{\tau_k})$ diagonal

matrix containing the non-zero singular values of \mathbf{A}_{τ_k} . Then,

$$\begin{aligned}
\mathbb{E}_{k-1} \|\mathbf{P}_{\tau_k} \mathbf{e}_{k-1}\|_2^2 &= \mathbb{E}_{k-1} \|\mathbf{U}_{\tau_k} \mathbf{U}_{\tau_k}^* \mathbf{e}_{k-1}\|_2^2 \\
&= \mathbb{E}_{k-1} \|\mathbf{U}_{\tau_k}^* \mathbf{e}_{k-1}\|_2^2 \\
&= \mathbb{E}_{k-1} \|\boldsymbol{\Sigma}_{\tau_k}^{-1} \mathbf{V}_{\tau_k}^* \mathbf{A}_{\tau_k}^* \mathbf{e}_{k-1}\|_2^2 \\
&\geq \mathbb{E}_{k-1} \sigma_{\min}^2(\boldsymbol{\Sigma}_{\tau_k}^{-1} \mathbf{V}_{\tau_k}^*) \|\mathbf{A}_{\tau_k}^* \mathbf{e}_{k-1}\|_2^2 \\
&= \mathbb{E}_{k-1} \frac{\|\mathbf{A}_{\tau_k}^* \mathbf{e}_{k-1}\|_2^2}{\sigma_{\max}^2(\boldsymbol{\Sigma}_{\tau_k})} \\
&\geq \frac{1}{\overline{p\beta}} \sum_{\tau_k \in \mathcal{T}} \|\mathbf{A}_{\tau_k}^* \mathbf{e}_{k-1}\|_2^2 \\
&= \frac{1}{\overline{p\beta}} \|\mathbf{A}^* \mathbf{e}_{k-1}\|_2^2 \\
&\geq \frac{\sigma_{\min}^2(\mathbf{A})}{\overline{p\beta}} \|\mathbf{e}_{k-1}\|_2^2,
\end{aligned}$$

where the first equality follows since $\mathbf{P}_{\tau_k} = \mathbf{U}_{\tau_k} \mathbf{U}_{\tau_k}^*$, the second equality from the unitary invariance property of the Euclidean norm, the third equality by replacing $\mathbf{U}_{\tau_k}^*$ with $\boldsymbol{\Sigma}_{\tau_k}^{-1} \mathbf{V}_{\tau_k}^* \mathbf{A}_{\tau_k}^*$, the next three lines follow since $\sigma_{\min}^2(\boldsymbol{\Sigma}_{\tau_k}^{-1} \mathbf{V}_{\tau_k}^*) = 1/\sigma_{\max}^2(\boldsymbol{\Sigma}_{\tau_k}) = 1/\sigma_{\max}^2(\mathbf{A}_{\tau_k})$ and by the paving assumption, and the final inequality follows since $\mathbf{e}_k \in \mathcal{R}(\mathbf{A})$ for all $k \geq 0$ (indeed, $\mathbf{e}_0 = \mathbf{b}_{\mathcal{R}(\mathbf{A})} \in \mathcal{R}(\mathbf{A})$ and it follows that $\mathbf{e}_k \in \mathcal{R}(\mathbf{A})$ for every $k \geq 0$ by the recursive definition of \mathbf{e}_k). It follows that

$$\mathbb{E}_{k-1} \|\mathbf{e}_k\|_2^2 = \|\mathbf{e}_{k-1}\|_2^2 - \mathbb{E}_{k-1} \|\mathbf{P}_{\tau_k} \mathbf{e}_{k-1}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\overline{p\beta}}\right) \|\mathbf{e}_{k-1}\|_2^2. \quad (8)$$

Repeat the above inequality k times and notice that $\mathbf{e}_0 = \mathbf{b} - \mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp} = \mathbf{b}_{\mathcal{R}(\mathbf{A})}$ to conclude. \square

Next, Lemma 2.2 of [NT13] shows that for any vector \mathbf{u} ,

$$\mathbb{E} \left\| \left(\mathbf{I} - (\mathbf{A}_\tau)^\dagger \mathbf{A}_\tau \right) \mathbf{u} \right\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{p\beta} \right) \|\mathbf{u}\|_2^2. \quad (9)$$

Since the range of $\mathbf{I} - (\mathbf{A}_\tau)^\dagger \mathbf{A}_\tau$ and $(\mathbf{A}_\tau)^\dagger$ are orthogonal, we have

$$\|\mathbf{x}_k - \mathbf{x}_{\text{LS}}\|_2^2 = \left\| \left(\mathbf{I} - (\mathbf{A}_\tau)^\dagger \mathbf{A}_\tau \right) (\mathbf{x}_{k-1} - \mathbf{x}_{\text{LS}}) \right\|_2^2 + \left\| (\mathbf{A}_\tau)^\dagger ((\mathbf{z}_k)_\tau - \mathbf{b}_\tau^\perp) \right\|_2^2, \quad (10)$$

where for shorthand we will write \mathbf{b}_τ^\perp to mean $(\mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp})_\tau$. Combining (9) with $\mathbf{u} = \mathbf{x}_{k-1} - \mathbf{x}_{\text{LS}}$

along with (10), we have

$$\begin{aligned}
\mathbb{E} \|\mathbf{x}_k - \mathbf{x}_{\text{LS}}\|_2^2 &\leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{p\beta}\right) \mathbb{E} \|\mathbf{x}_{k-1} - \mathbf{x}_{\text{LS}}\|_2^2 + \mathbb{E} \left\| (\mathbf{A}_\tau)^\dagger ((\mathbf{z}_k)_\tau - \mathbf{b}_\tau^\perp) \right\|_2^2 \\
&\leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{p\beta}\right) \mathbb{E} \|\mathbf{x}_{k-1} - \mathbf{x}_{\text{LS}}\|_2^2 + \frac{1}{\sigma_{\min}^2(\mathbf{A}_\tau)} \mathbb{E} \|(\mathbf{z}_k)_\tau - \mathbf{b}_\tau^\perp\|_2^2 \\
&\leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{p\beta}\right) \mathbb{E} \|\mathbf{x}_{k-1} - \mathbf{x}_{\text{LS}}\|_2^2 + \frac{1}{\alpha} \mathbb{E} \|\mathbf{z}_k - \mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp}\|_2^2.
\end{aligned} \tag{11}$$

To apply this bound recursively, we will utilize an elementary lemma. It is essentially proved in [ZF12, Theorem 8] but for completeness we recall its proof in the appendix.

Lemma 3. *Suppose that for some $\gamma, \bar{\gamma} < 1$, the following bounds hold for all $k^* \geq 0$:*

$$\mathbb{E} \|\mathbf{x}_{k^*} - \mathbf{x}_{\text{LS}}\|_2^2 \leq \gamma \mathbb{E} \|\mathbf{x}_{k^*-1} - \mathbf{x}_{\text{LS}}\|_2^2 + r_{k^*} \quad \text{and} \quad r_{k^*} \leq \bar{\gamma}^{k^*} B. \tag{12}$$

Then for any $T > 0$,

$$\mathbb{E} \|\mathbf{x}_T - \mathbf{x}_{\text{LS}}\|_2^2 \leq \gamma^T \|\mathbf{x}_0 - \mathbf{x}_{\text{LS}}\|_2^2 + \left(\gamma^{\lfloor T/2 \rfloor} + \bar{\gamma}^{\lfloor T/2 \rfloor} \right) \frac{B}{1 - \gamma}.$$

Using $r_k = \frac{1}{\alpha} \mathbb{E} \|\mathbf{z}_k - \mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp}\|_2^2$, $B = \frac{\|\mathbf{b}_{\mathcal{R}(\mathbf{A})}\|_2^2}{\alpha}$, $\gamma = 1 - \frac{\sigma_{\min}^2(\mathbf{A})}{p\beta}$, and $\bar{\gamma} = 1 - \frac{\sigma_{\min}^2(\mathbf{A})}{p\beta}$, we see by Lemma 2 and (11) that the bounds (12) hold. Applying Lemma 3 completes the proof. \square

2.1. Implementation

In Section 3 we will see that matrix row-pavings for *standardized* matrices, those whose rows have unit norm, can be obtained readily. One can thus use the column-normalized version of the matrix \mathbf{A} in line 5 of Algorithm 1 and the corresponding column paving. Note that one need not have access to the complete column-standardized matrix, instead the columns of the submatrix in line 5 could be normalized on the fly. See Section 3 for more on obtaining row pavings and further details.

2.2. Comparison of Convergence Rates

The bound of Theorem 1 improves upon that of the randomized block Kaczmarz method because it demonstrates linear convergence to the least squares solution \mathbf{x}_{LS} , whereas (6) shows convergence only within a radius proportional to $\|\mathbf{e}\|_2^2$, which we call the *convergence horizon*. Algorithm 1 is able to break this barrier because it iteratively removes the component of \mathbf{b} which is orthogonal to the range of \mathbf{A} . This of course is also true of the randomized Extended Kaczmarz method (3) as it also breaks this horizon barrier. To compare the rate of (4) to that of Theorem 5, we consider two important scenarios.

First, consider the case when A is nearly square, and each submatrix can be applied efficiently via a fast multiply. In this case, each iteration of Algorithm 1 incurs approximately the same computational cost as an iteration of the REK method. Thus, we may directly compare the convergence rates of Theorem 5 and (4) to find that Algorithm 1 is about $n/(p\beta)$ times faster than REK in this setting. Thus when n is much larger than $p\beta$, this can result in a significant speedup.

Alternatively, if the matrix A does not admit a fast multiply, it is fair to only compare the convergence rate per epoch, since each iteration of Algorithm 1 may require more computational cost than those of REK. Since an epoch of Algorithm 1 and REK consist of p and n iterations, respectively, we see that the rate of the former is proportional to $\sigma_{\min}^2(A)/\beta$ whereas that of REK is proportional to $\sigma_{\min}^2(A)$. We see in this case that these bounds suggest REK exhibits faster convergence (assuming $\beta > 1$). However, as observed in the randomized Block Kaczmarz method, the block methods still display faster convergence than their single counterparts because of implicit computational issues in the linear algebraic subroutines. See the discussion in [NT13] and the experimental results below for further details.

3. Obtaining matrix pavings

We devote this section to a brief discussion about matrix pavings and how they may be obtained to utilize the results of Theorem 1. The results discussed here on matrix pavings stem from *subset selection*, the problem of selecting a large submatrix with desired geometric properties, whose origins come from the well-known Restricted Invertibility Principle of Bourgain Tzafriri [BT87]. The literature now contains several results on subset selection, see e.g. [BT91, KT94, Ver01, Ver06, Tro09b] and [Sri10, Nao11, SS12, You12b] for recent advancements. We summarize here a few results useful for our purposes.

The simplest type of matrix to first consider is one which has unit-norm rows, which we call *row-standardized* (and a matrix with unit-norm columns is *column-standardized*). A surprising result shows that every row-standardized matrix admits a row paving with well-controlled paving parameters. Tropp proves the following result in [Tro09a, Thm. 1.2], whose origins are due to Bourgain and Tzafriri [BT87, BT91] and Vershynin [Ver06].

Proposition 4 (Existence of Good Pavings). *Fix a number $\delta \in (0, 1)$ and row-standardized matrix A with n rows. Then A admits a (p, α, β) row paving with*

$$p \leq C \cdot \delta^{-2} \|A\|^2 \log(1+n) \quad \text{and} \quad 1 - \delta \leq \alpha \leq \beta \leq 1 + \delta, \quad (13)$$

where C denotes an absolute constant.

Proposition 4 shows the *existence* of such a paving, but the literature provides various efficient mechanisms for the construction of good pavings as well. In many cases one constructs such a paving simply by choosing a partition of an appropriate size *at random*, see [Tro09a] for an efficient method to compute a paving satisfying (13). See also [NT13] and the references therein for a thorough discussion of these types of results.

If the matrix \mathbf{A} is row-standardized, one can thus construct a row paving satisfying (13). If the matrix also naturally admits a column paving (for example if it is symmetric or positive semi-definite), then both pavings will have such bounded parameters. If the latter does not hold, one can instead utilize the column-standardized version of \mathbf{A} , which we denote $\bar{\mathbf{A}}$, in line 5 of Algorithm 1. The standardization can either be done on the fly during the algorithm, or ahead of time. Either way, Proposition 4 can be combined with Theorem 1 to yield the following corollary.

Corollary 5. *Suppose Algorithm 1 is run on a row-standardized matrix \mathbf{A} , \mathbf{b} , $T \in \mathbb{N}$, $(\bar{p}, \bar{\alpha}, \bar{\beta})$ column paving \mathcal{T} of the column-standardized version $\bar{\mathbf{A}}$, and (p, α, β) row paving \mathcal{S} of \mathbf{A} , both guaranteed by Proposition 4. Then the estimate vector \mathbf{x}_T satisfies*

$$\mathbb{E} \|\mathbf{x}_T - \mathbf{x}_{LS}\|_2^2 \leq \gamma^T \|\mathbf{x}_0 - \mathbf{x}_{LS}\|_2^2 + \left(\gamma^{\lfloor T/2 \rfloor} + \bar{\gamma}^{\lfloor T/2 \rfloor} \right) \frac{C \|\mathbf{b}_{\mathcal{R}(\mathbf{A})}\|_2^2}{(1 - \gamma)},$$

where $\gamma = 1 - \frac{C}{\kappa^2(\mathbf{A}) \log(1+n)}$, $\bar{\gamma} = 1 - \frac{C}{\kappa^2(\bar{\mathbf{A}}) \log(1+d)}$ and $\kappa(\mathbf{A}) = \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}$ and $\kappa(\bar{\mathbf{A}}) = \frac{\sigma_{\max}(\bar{\mathbf{A}})}{\sigma_{\min}(\bar{\mathbf{A}})}$ denote the condition numbers of \mathbf{A} and $\bar{\mathbf{A}}$, respectively.

If the matrix \mathbf{A} is not row-standardized, one can run Algorithm 1 on the standardized system, along with a paving guaranteed by Proposition 4. Clearly, the method will converge to the new least squares solution, which does not necessarily coincide with the original least squares solution in the inconsistent case.

Alternatively, if the matrix \mathbf{A} is not row-standardized and one wishes to ensure convergence to the true least-squares solution \mathbf{x}_{LS} , Proposition 4 can still be used with sub-optimal paving bounds. Indeed, write $\tilde{\mathbf{D}}$ as the diagonal matrix whose entries correspond to the reciprocals of the row-norms $\|\mathbf{a}_i\|_2$ of \mathbf{A} , so that $\tilde{\mathbf{A}} = \tilde{\mathbf{D}}\mathbf{A}$ has unit-norm rows. Set

$$a_{\max} = \max_i \|\mathbf{a}_i\|_2^2 \quad \text{and} \quad a_{\min} = \min_i \|\mathbf{a}_i\|_2^2$$

Utilize Proposition 4 on $\tilde{\mathbf{A}}$ to obtain a row-paving with parameters \tilde{p} , $\tilde{\alpha}$ and $\tilde{\beta}$ so that

$$\tilde{p} \leq C \cdot \delta^{-2} \|\tilde{\mathbf{A}}\|^2 \log(1+n) \quad \text{and} \quad 1 - \delta \leq \tilde{\alpha} \leq \tilde{\beta} \leq 1 + \delta.$$

If one uses this same paving for \mathbf{A} , one has (quite pessimistically) that the corresponding paving parameters p , α , and β for \mathbf{A} satisfy

$$p \leq C \cdot \delta^{-2} a_{\min}^{-1} \|\mathbf{A}\|^2 \log(1+n) \quad \text{and} \quad a_{\min}(1 - \delta) \leq \alpha \leq \beta \leq a_{\max}(1 + \delta).$$

One can then directly apply Theorem 1 with this paving to obtain an analogous version of Corollary 5 for arbitrary matrices, which shows that

$$\mathbb{E} \|\mathbf{x}_T - \mathbf{x}_{LS}\|_2^2 \leq \gamma^T \|\mathbf{x}_0 - \mathbf{x}_{LS}\|_2^2 + \left(\gamma^{\lfloor T/2 \rfloor} + \bar{\gamma}^{\lfloor T/2 \rfloor} \right) \frac{C \|\mathbf{b}_{\mathcal{R}(\mathbf{A})}\|_2^2}{a_{\min}(1 - \gamma)}, \quad (14)$$

where now $\gamma = 1 - \frac{C a_{\min}}{a_{\max} \kappa^2(A) \log(1+n)} = 1 - \frac{C}{a_r \kappa^2(A) \log(1+n)}$ where we have written $a_r = a_{\max}/a_{\min}$ to denote the dynamic range of the row norms of A (and $\bar{\gamma}$ remains the same as in the corollary). This demonstrates that if the dynamic range is bounded, the convergence rate is the same as in the standardized case, up to constants. In either case, these results demonstrate linear convergence to the true least squares solution.

An alternative to these types of bounds can be obtained by simply only using a column paving for the matrix. As we elaborate below, the least squares solution can still be obtained from the column-normalized system. Since column-pavings can be easily attained in this case via Proposition 4, such a method offers a nice alternative in situations where the dynamic range of the row norms is unknown or unbounded. We propose such a method in the next section.

4. A Randomized Block Coordinate Descent Method

We next present a simple variant of the extended Kaczmarz method which utilizes only a column paving of the matrix. For that reason, one need not worry about whether the matrix A is row-standardized. Moreover, the column-standardized version can be used within the algorithm which guarantees bounded paving parameters, while still finding the true least squares solution.

Utilizing the benefits of both the block variant and the randomized extension, we propose the following randomized block coordinate descent method for the inconsistent case.

Algorithm 2 Randomized Block Least Squares Solver

```

1: procedure ( $A, \mathbf{b}, T, \mathcal{T}$ )  $\triangleright A \in \mathbb{R}^{n \times d}, \mathbf{b} \in \mathbb{R}^n, T \in \mathbb{N}$ , column partition  $\mathcal{T}$  of  $[d]$ 
2:   Initialize  $\mathbf{x}_0 = \mathbf{0}$  and  $\mathbf{z}_0 = \mathbf{b}$ 
3:   for  $k = 1, 2, \dots, T$  do
4:     Pick  $\tau_k \in \mathcal{T}$  uniformly at random
5:     Compute  $\mathbf{w}_k = (A_{\tau_k})^\dagger \mathbf{z}_{k-1}$   $\triangleright A_{\tau_k}$ :  $n \times |\tau_k|$  submatrix of  $A$ 
6:     Update  $(\mathbf{x}_k)_{\tau_k} = (\mathbf{x}_{k-1})_{\tau_k} + \mathbf{w}_k$ 
7:     Set  $\mathbf{z}_k = \mathbf{z}_{k-1} - A_{\tau_k} \mathbf{w}_k$ 
8:   end for
9:   Output  $\mathbf{x}_T$ 
10: end procedure

```

4.1. Analysis of Randomized Block Least Squares Solver

We may utilize some of the previous analysis to prove convergence of Algorithm 2. Observe that Step 6 of Algorithm 1 is identical to Steps 5 and 7 of Algorithm 2, therefore Lemma 2 implies that

$$\mathbb{E} \|\mathbf{z}_k - \mathbf{b}_{\mathcal{R}(A)^\perp}\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(A)}{\bar{p}\bar{\beta}}\right)^k \|\mathbf{b}_{\mathcal{R}(A)}\|_2^2. \quad (15)$$

To utilize this result, we aim to relate the iterates \mathbf{z}_k to the estimation \mathbf{x}_k . The following claim quantifies precisely this relation.

Lemma 6. *For every $k \geq 0$, at the end of the k -th iteration, it holds that $\mathbf{z}_{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}_{k+1}$.*

Proof. We prove by induction on k . Instate the notation of Algorithm 2, and for two sets S_1 and S_2 write $S_1 \setminus S_2 = S_1 \cap S_2^c$ to denote set subtraction. For the base case of $k = 0$, we have $(\mathbf{x}_1)_{\tau_0} = \mathbf{w}_0$ and $(\mathbf{x}_1)_{[n] \setminus \tau_0} = \mathbf{0}$ and moreover, $\mathbf{z}_1 = \mathbf{z}_0 - \mathbf{A}_{\tau_0} \mathbf{w}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_1$. Assume that $\mathbf{z}_\ell = \mathbf{b} - \mathbf{A}\mathbf{x}_\ell$ is true for some $\ell > 0$, we will show that it holds for $\ell + 1$. For the sake of notation, denote $\mathbf{P}_\ell = \mathbf{A}_{\tau_\ell} \mathbf{A}_{\tau_\ell}^\dagger$. Then

$$\mathbf{z}_{\ell+1} = \mathbf{z}_\ell - \mathbf{P}_\ell \mathbf{z}_\ell = \mathbf{b} - \mathbf{A}\mathbf{x}_\ell - \mathbf{P}_\ell \mathbf{z}_\ell \quad (16)$$

the first equality follows by the definition of $\mathbf{z}_{\ell+1}$, the second equality follows by induction hypothesis. Now, it follows that

$$\begin{aligned} \mathbf{A}\mathbf{x}_{\ell+1} &= \mathbf{A}_{\tau_\ell} (\mathbf{x}_{\ell+1})_{\tau_\ell} + \mathbf{A}_{[n] \setminus \tau_\ell} (\mathbf{x}_{\ell+1})_{[n] \setminus \tau_\ell} \\ &= \mathbf{A}_{\tau_\ell} (\mathbf{x}_\ell)_{\tau_\ell} + \mathbf{A}_{\tau_\ell} \mathbf{w}_\ell + \mathbf{A}_{[n] \setminus \tau_\ell} (\mathbf{x}_{\ell+1})_{[n] \setminus \tau_\ell} \\ &= \mathbf{A}_{\tau_\ell} (\mathbf{x}_\ell)_{\tau_\ell} + \mathbf{A}_{\tau_\ell} \mathbf{w}_\ell + \mathbf{A}_{[n] \setminus \tau_\ell} (\mathbf{x}_\ell)_{[n] \setminus \tau_\ell} \\ &= \mathbf{A}\mathbf{x}_\ell + \mathbf{A}_{\tau_\ell} \mathbf{w}_\ell. \end{aligned}$$

the first equality follows by Step 6 of the algorithm (update on \mathbf{x}), the second equality because $(\mathbf{x}_{\ell+1})_{[n] \setminus \tau_\ell} = (\mathbf{x}_\ell)_{[n] \setminus \tau_\ell}$. Hence, $\mathbf{A}\mathbf{x}_\ell = \mathbf{A}\mathbf{x}_{\ell+1} - \mathbf{A}_{\tau_\ell} \mathbf{w}_\ell$. Now, the right hand side of (16) can be rewritten as

$$\begin{aligned} \mathbf{b} - \mathbf{A}\mathbf{x}_\ell - \mathbf{P}_\ell \mathbf{z}_\ell &= \mathbf{b} - \mathbf{A}\mathbf{x}_{\ell+1} + \mathbf{A}_{\tau_\ell} \mathbf{w}_\ell - \mathbf{P}_\ell \mathbf{z}_\ell \\ &= \mathbf{b} - \mathbf{A}\mathbf{x}_{\ell+1}. \end{aligned}$$

the last equality follows since $\mathbf{w}_\ell = (\mathbf{A}_{\tau_\ell})^\dagger \mathbf{z}_\ell$. Therefore, we conclude that $\mathbf{z}_{\ell+1} = \mathbf{b} - \mathbf{A}\mathbf{x}_{\ell+1}$ which completes the proof. \square

Combining this lemma with (15) yields the following result which shows convergence of the estimation to the least squares solution under the map \mathbf{A} .

Theorem 7. *Algorithm 2 with input \mathbf{A} , \mathbf{b} , $T \in \mathbb{N}$, and $(\bar{p}, \bar{\alpha}, \bar{\beta})$ column paving \mathcal{T} , outputs an estimate vector \mathbf{x}_T that satisfies*

$$\mathbb{E} \|\mathbf{A}(\mathbf{x}_{\text{LS}} - \mathbf{x}_T)\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\bar{p}\bar{\beta}}\right)^T \|\mathbf{b}_{\mathcal{R}(\mathbf{A})}\|_2^2.$$

Proof. We observe that

$$\mathbf{A}(\mathbf{x}_{\text{LS}} - \mathbf{x}_{(k)}) = \mathbf{b}_{\mathcal{R}(\mathbf{A})} - \mathbf{A}\mathbf{x}_{(k)} = \mathbf{b} - \mathbf{A}\mathbf{x}_{(k)} - \mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp} = \mathbf{z}_{(k)} - \mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp}$$

where the first equality follows by $\mathbf{b}_{\mathcal{R}(\mathbf{A})} = \mathbf{A}\mathbf{A}^\dagger \mathbf{b} = \mathbf{A}\mathbf{x}_{\text{LS}}$, the second by orthogonality $\mathbf{b} = \mathbf{b}_{\mathcal{R}(\mathbf{A})^\perp} + \mathbf{b}_{\mathcal{R}(\mathbf{A})}$ and the last equality from Lemma 6. Combined with inequality (7) this yields the desired result. \square

When \mathbf{A} has full column rank, we may bound the estimation error $\|\mathbf{x}_{\text{LS}} - \mathbf{x}_T\|_2$ by $\frac{1}{\sigma_{\min}(\mathbf{A})} \|\mathbf{A}(\mathbf{x}_{\text{LS}} - \mathbf{x}_T)\|_2$ which combined with the fact that $\|\mathbf{b}_{\mathcal{R}(\mathbf{A})}\|_2 \leq \sigma_{\max}(\mathbf{A}) \|\mathbf{x}_{\text{LS}}\|_2$ implies the following corollary.

Corollary 8. *Algorithm 2 with full-rank \mathbf{A} , \mathbf{b} , $T \in \mathbb{N}$, and $(\bar{p}, \bar{\alpha}, \bar{\beta})$ column paving \mathcal{T} , outputs an estimate vector \mathbf{x}_T that satisfies*

$$\mathbb{E} \|\mathbf{x}_{\text{LS}} - \mathbf{x}_T\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\mathbf{A})}{\bar{p}\bar{\beta}}\right)^T \kappa^2(\mathbf{A}) \|\mathbf{x}_{\text{LS}}\|_2^2.$$

4.2. Implementation

The advantage to a single paving approach as in Algorithm 2 is that one can utilize the column-standardized version of \mathbf{A} while maintaining the same convergence to the (scaled version of the) least squares solution. Utilizing Proposition 4, one is guaranteed a column-paving satisfying (13), so that paving parameters of Theorem 7 and Corollary 8 are bounded. Since re-normalizing the columns of the matrix only re-scales the entries of \mathbf{x}_{LS} , Theorem 7 and Corollary 8 grant one access to the original least squares solution \mathbf{x}_{LS} . To be precise, now let \mathbf{D} be the diagonal matrix whose entries correspond to the reciprocals of the column norms of \mathbf{A} , so that $\bar{\mathbf{A}} = \mathbf{A}\mathbf{D}$ has unit-norm columns. Let $\bar{\mathbf{x}}_{\text{LS}} = \bar{\mathbf{A}}^\dagger \mathbf{b}$ denote the least squares solution of the re-normalized system, so that one has $\bar{\mathbf{x}}_{\text{LS}} = \mathbf{D}^{-1} \mathbf{x}_{\text{LS}}$. Then

$$\|\bar{\mathbf{A}}(\bar{\mathbf{x}}_{\text{LS}} - \mathbf{x}_T)\|_2^2 = \|\mathbf{A}\mathbf{D}(\mathbf{D}^{-1} \mathbf{x}_{\text{LS}} - \mathbf{x}_T)\|_2^2 = \|\mathbf{A}(\mathbf{x}_{\text{LS}} - \mathbf{D}\mathbf{x}_T)\|_2^2.$$

Thus applying Theorem 7 for $\bar{\mathbf{A}}$, and utilizing the fact that the range of \mathbf{A} is the same as that of $\bar{\mathbf{A}}$, one has

$$\mathbb{E} \|\mathbf{A}(\mathbf{x}_{\text{LS}} - \mathbf{D}\mathbf{x}_T)\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\bar{\mathbf{A}})}{\bar{p}\bar{\beta}}\right)^T \|\mathbf{b}_{\mathcal{R}(\mathbf{A})}\|_2^2.$$

This then implies that when the matrix is full rank,

$$\mathbb{E} \|\mathbf{x}_{\text{LS}} - \mathbf{D}\mathbf{x}_T\|_2^2 \leq \left(1 - \frac{\sigma_{\min}^2(\bar{\mathbf{A}})}{\bar{p}\bar{\beta}}\right)^T \kappa^2(\mathbf{A}) \|\mathbf{x}_{\text{LS}}\|_2^2. \quad (17)$$

Since \bar{p} and $\bar{\beta}$ are the paving parameters of $\bar{\mathbf{A}}$, one has by utilizing Proposition 4 and substituting the bounds of (13) into (17) that

$$\mathbb{E} \|\mathbf{x}_{\text{LS}} - \mathbf{D}\mathbf{x}_T\|_2^2 \leq \left(1 - \frac{C'}{\kappa^2(\bar{\mathbf{A}}) \log(1+d)}\right)^T \kappa^2(\mathbf{A}) \|\mathbf{x}_{\text{LS}}\|_2^2. \quad (18)$$

Although this bound does depend on the conditioning of both \mathbf{A} and $\bar{\mathbf{A}}$ (note the rate itself only depends on the conditioning of the latter), it is the first to guarantee linear convergence to the true least squares solution utilizing a paving which is guaranteed by Proposition 4 while not placing any restrictions on the matrix \mathbf{A} itself (such as standardization).

Remark 1. *In considering the improvements offered by both the REK method and the block Kaczmarz method, one may ask whether it is advantageous to run a traditional REK projection step as in (3) along with a traditional block Kaczmarz update step as in (5). However, empirically we have observed that such a combination actually leads to a degradation in performance and requires far more epochs to converge than the algorithms discussed above. We conjecture that it is important to run both the projection update and the Kaczmarz update “at the same speed”; if the Kaczmarz update utilizes many rows at once, so should the projection update, and vice versa.*

5. Summary of Approaches

Here we detail the various approaches proposed, and summarize the practical implementation in several frameworks. Since the block variants of the methods require the blocks be well conditioned, it is natural to rely on matrix pavings for the analysis. Unfortunately, such pavings are only readily available when the matrix itself is properly normalized. For that reason, we have proposed several practical alternatives for the important setting in which normalization is not feasible. Both of our proposed methods still offer computational advantages over the standard approaches (see the next section). We summarize these approaches here, which cover all possible settings.

Consistent systems: When the system is consistent, one does not lose any convergence properties by re-scaling the system to be standardized, since the solution remains the same. In this simple setting, one can utilize the standardized system (either standardizing a priori or on the fly), and benefit from the convergence guaranteed by Corollary 5.

Inconsistent standardized systems: When the matrix is already standardized (as naturally occurs for example in Vandermonde matrices used in trigonometric approximation), as in the above case one can immediately utilize Corollary 5 to guarantee convergence to the least squares solution.

Inconsistent systems with bounded dynamic range: If the system is not standardized, but the ratio of the largest to smallest row norm is bounded (as is the case in random matrices for example, which have tightly concentrated row norms), one can still utilize this result. Indeed, by utilizing the same paving one would use if the matrix was actually standardized – but not actually standardizing the system, Corollary 5 can be used to obtain the convergence rate given in (14). One sees that if the dynamic range a_r is bounded (by say, a constant or even $\log n$), the method converges to the original least squares solution with approximately the same convergence rate.

Inconsistent systems with unbounded dynamic range: If the matrix has a large variety of row norms, it may clearly be challenging to guarantee the desired row paving. For that reason, it will be advantageous to use a column paving. To that end, we propose Algorithm 2 which is designed for systems that cannot be separated into well-conditioned row blocks. As discussed in Section 4.2, Theorem 7 can be used

via the column-standardized paving to obtain the convergence rate given in (18). This bound guarantees linear convergence to the true least squares solution, even for matrices which are far from standardized. The disadvantage of course is that the rate depends on the conditioning of the column-standardized version, which may be hard to explicitly bound (as is true for general large matrices anyway).

In any of these cases, our results may be used to guarantee linear convergence in expectation to the true least squares solution of the system. Moreover, because matrix blocks can be utilized, we often see a significant speedup in runtime due to practical considerations. See the next section for examples of such behavior.

6. Experimental Results

Here we present some experiments using simple examples to illustrate the benefits of block methods. We do not claim optimized implementations of the method, and only run on small problem sizes; our purpose is only to demonstrate that even in these simple examples, the block method offers advantages to the standard method. We refer the reader to [ZF12, NT13] for more empirical results for both REK and block methods.

In all experiments, one matrix is created and 40 trials of each method are run. In our first experiment, the matrix is a 300×100 matrix with standard normal entries, whose rows are then normalized to each have norm one, yielding a condition number of 3.7. We use 30 blocks, selected by a random partition. The vector \mathbf{x} is created to have independent standard normal entries, and the right hand side \mathbf{b} is set to \mathbf{Ax} . We track the ℓ_2 -error $\|\mathbf{x}_{LS} - \mathbf{x}_k\|_2$ across each epoch³ as well as the CPU time (measured in Matlab using the `cputime` command). In all experiments we considered a trial successful when the error reached 10^{-6} . The results for this case are presented in Figures 1 and 2. In all figures, a heavy line represents median performance, and the shaded region spans the minimum to the maximum value across all trials. As is demonstrated, even when the matrix does not have any natural block structure, the proposed algorithms outperform standard REK in terms of runtime.

Figure 3 shows similar plots, but in this case the system is no longer consistent. For these experiments, we used the same type and size of the matrix \mathbf{A} , but the right hand side vector \mathbf{b} was generated as a Gaussian vector as well. We created \mathbf{b} so that the residual norm $\|\mathbf{b} - \mathbf{Ax}_{LS}\|_2 = 0.5$. We then track the ℓ_2 -error between the iterate \mathbf{x}_k and the least squares solution \mathbf{x}_{LS} which we computed by $\mathbf{A}^\dagger \mathbf{b}$. We repeat the experiment also for a matrix whose dynamic range is not well bounded. For that experiment, we generate a Gaussian matrix and then scale the row norms so that the i th row has norm equal to i .

³We refer to an epoch as the number of iterations that is equivalent to one cycle through n rows, even though rows and blocks are selected with replacement. Thus for REK, an epoch is n iterations, and for a block version with b blocks, one epoch is b iterations. For an attempt at a fair comparison with Algorithm 2 that only uses a column paving, we measure an “epoch” to be n/b where b is the number of blocks in the column paving.

The results for this case using Algorithm 2 are shown in Figure 4. The behavior in both cases, as predicted by our main results, is quite similar to the consistent case and thus breaks the convergence horizon of the standard Kaczmarz method.

Lastly, we tested the methods on tomography problems, generated using the Matlab Regularization Toolbox by P.C. Hansen (<http://www.imm.dtu.dk/~pcha/Regutools/>) [Han07]. In particular we present a 2D tomography problem $\mathbf{Ax} = \mathbf{b}$ for an $n \times d$ matrix with $n = fN^2$ and $d = N^2$. Here \mathbf{A} corresponds to the absorption along a random line through an $N \times N$ grid. In our experiments we set $N = 20$ and the oversampling factor $f = 3$. This yielded a matrix \mathbf{A} with condition number $\kappa(\mathbf{A}) = 2.08$. Since for this matrix it may be difficult to obtain a row paving, we instead use Algorithm 2 and obtain a random column paving from the standardized version and use that for the matrix \mathbf{A} . The results for various choices of paving size are displayed in Figure 5, which are in line with previous experiments.

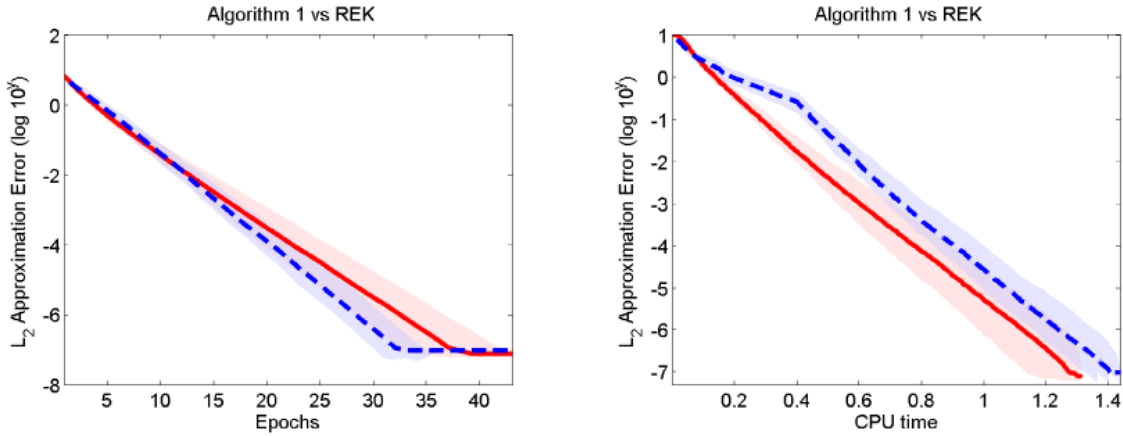


Figure 1: ℓ_2 -norm error for REK (blue dashed) and Algorithm 1 (red) across epochs (left) and CPU time (right). Matrix is 300×100 Gaussian, system is consistent.

7. Related Work and Discussion

The Kaczmarz method was first introduced in the 1937 work of Kaczmarz himself [Kac37]. Since then, the method has been revitalized by researchers in computer tomography, under the name *Algebraic Reconstruction Technique* (ART) [GBH70, Byr08, Nat01, Her09]. Deterministic convergence results for the method often depend on properties of the matrix that are difficult to compute or analyze [Deu85, DH97, XZ02, Gal05]. Moreover, it has been well observed that random choice of row selection often speeds up the convergence [HS78, HM93, CFM⁺92, Nat01].

Recently, Strohmer and Vershynin [SV09] derived the first provable convergence rate of the Kaczmarz method, showing that when each row is selected with probability proportional to its norm the method exhibits the expected linear convergence of (1). This work was extended to the inconsistent case in [Nee10], which shows linear convergence

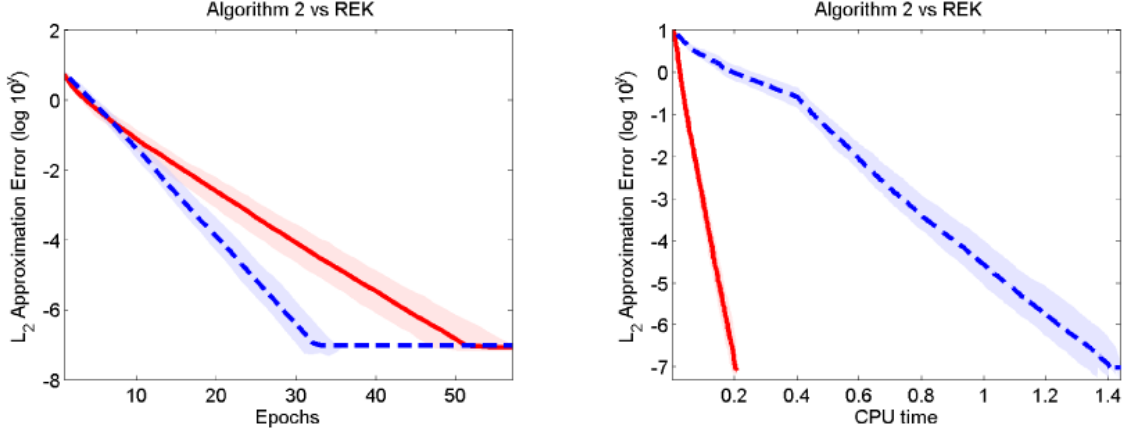


Figure 2: ℓ_2 -norm error for REK (blue dashed) and Algorithm 2 (red) across epochs (left) and CPU time (right). Matrix is 300×100 Gaussian, system is consistent.

to within some fixed radius of the least squares solution. The almost-sure guarantees were recently derived by Chen and Powell [CP12]. To break the convergence barrier, relaxation parameters can be introduced, so that each iterate is over or under projected onto each solution space. Whitney and Meany prove that if the relaxation parameters tend to zero that the iterates converge to the least squares solution [WM67]. Further results using relaxation have also been obtained, see for example [CEG83, Tan71, HN90, ZF12]. An alternative to relaxation parameters was recently proposed by Zouzias and Freris [ZF12] as the REK method described by (3). Rather than alter the projection step, motivated by ideas of Popa [Pop98] they introduce a secondary step which aims to reduce the residual.

The Kaczmarz method has been extended beyond linear systems as well. For example, Leventhal and Lewis [LL10] analyze the method for systems with polyhedral constraints and inequalities, which was also extended to the block case [BN], and Richtárik and Takáč [RT11] build on these results for general optimization problems.

Another important aspect of research in this area focuses on accelerating the convergence of the methods. Geometric brute force methods can be used [EN11], additional row directions may be added [PPKR12], or instead one can select *blocks* of rows rather than a single row in each iteration. The block version of the Kaczmarz method is originally due to work of Elfving [Elf80] and Eggermont et al. [EHL81]. Its convergence rates were recently studied in [NW13] and analyzed via pavings by Needell and Tropp [NT13]. The block Kaczmarz method is of course a special instance in a broader class of block projection algorithms, see for example [XZ02] for a more general analysis and [Byr08] for a presentation of other block variants.

To use block methods effectively, one needs to obtain a suitable partition of the rows (and/or columns). Popa constructs such partitions by creating orthogonal blocks [Pop99, Pop01, Pop04], whereas Needell and Tropp promote the use of row pavings to construct the partition [NT13].

Construction of pavings has been studied for quite some time now, and most early

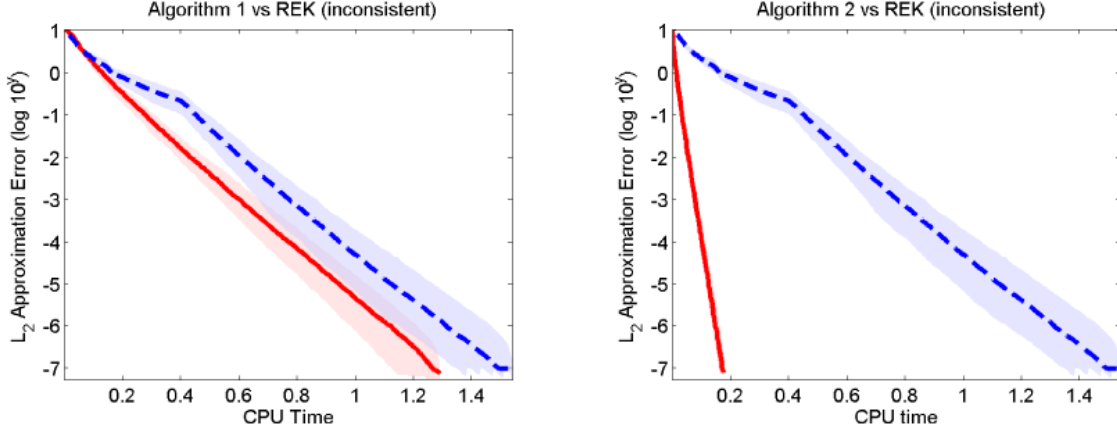


Figure 3: Matrix is 300×100 Gaussian, system is inconsistent. Left: ℓ_2 -norm error for REK (blue dashed) and Algorithm 1 (red) versus CPU time. Right: ℓ_2 -norm error for REK (blue dashed) and Algorithm 2 (red) versus CPU time.

results rely on random selection. The guarantee of lower and upper paving bounds has been derived by Bourgain and Tzafriri [BT87] and Kashin and Tzafriri [KT94], respectively. Simultaneous guarantees were later derived by Bourgain and Tzafriri [BT91] with suboptimal dependence on the matrix norm. Recently, Spielman and Srivastava [SS12] and Youssef [You12b] provided simple proofs of the results from [BT87] and [KT94], respectively. Vershynin [Ver01] and Srivastava [Sri10] extend the paving results to general matrices with arbitrary row norms; see also [You12b, You12a]. Proposition 4 follows from the work of Vershynin [Ver06] and Tropp [Tro09a], and is attributed to the seminal work of Bourgain and Tzafriri [BT87, BT91]. For particular classes of matrices, the paving can even be obtained from a random partition of the rows with high probability. This is proved by Tropp [Tro08a] using ideas from [BT91, Tro08b], and is refined in [CD12].

Acknowledgments

D.N. is thankful to the Simons Foundation Collaboration grant and the Alfred P. Sloan Fellowship. A.Z. has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC grant agreement n^o 259569. We would also like to thank Anna Ma for thoughtful discussions, and the reviewers for useful comments which significantly improved the manuscript.

Appendix A. Proof of intermediate results

Here we include the proof of Lemma 3.

Proof. [Proof of Lemma 3]

Assume the bounds (12) hold. Applying the first bound in (12) recursively yields

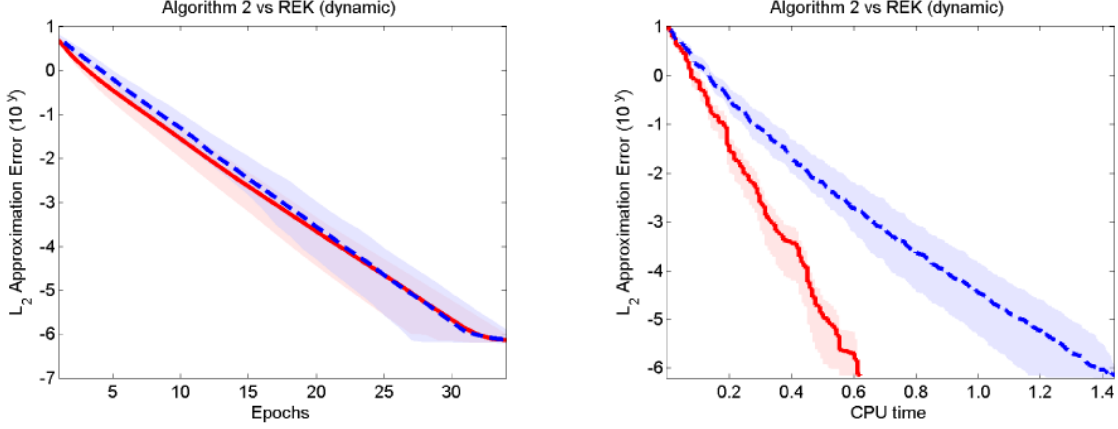


Figure 4: Matrix is 300×100 Gaussian with dynamic row norms, system is inconsistent. Plot shows ℓ_2 -norm error for REK (blue dashed) and Algorithm 2 (red) versus epochs (left) and CPU time (right).

$$\begin{aligned}
\mathbb{E} \|\mathbf{x}_{k^*} - \mathbf{x}_{\text{LS}}\|_2^2 &\leq \gamma^{k^*} \|\mathbf{x}_0 - \mathbf{x}_{\text{LS}}\|_2^2 + \sum_{j=0}^{k^*-1} \gamma^{k^*-1-j} r_j \\
&\leq \gamma^{k^*} \|\mathbf{x}_0 - \mathbf{x}_{\text{LS}}\|_2^2 + \sum_{j=0}^{\infty} \gamma^j B \\
&\leq \gamma^{k^*} \|\mathbf{x}_0 - \mathbf{x}_{\text{LS}}\|_2^2 + \frac{B}{1-\gamma},
\end{aligned}$$

where the second inequality holds by the assumption that $r_k \leq \gamma^k B \leq B$, and the last by the properties of the geometric summation. Similarly, observe that for any k and k^* we have

$$\begin{aligned}
\mathbb{E} \|\mathbf{x}_{k+k^*} - \mathbf{x}_{\text{LS}}\|_2^2 &\leq \gamma^k \mathbb{E} \|\mathbf{x}_{k^*} - \mathbf{x}_{\text{LS}}\|_2^2 + \sum_{j=0}^{k-1} \gamma^{k-1-j} r_{j+k^*} \\
&\leq \gamma^k \mathbb{E} \|\mathbf{x}_{k^*} - \mathbf{x}_{\text{LS}}\|_2^2 + \bar{\gamma}^{k^*} \sum_{j=0}^{\infty} \gamma^j B \\
&\leq \gamma^k \mathbb{E} \|\mathbf{x}_{k^*} - \mathbf{x}_{\text{LS}}\|_2^2 + \bar{\gamma}^{k^*} \frac{B}{1-\gamma}.
\end{aligned}$$

Now we choose k and k^* such that $T = k + k^*$ and $k = k^*$ if T is even, or $k = k^* + 1$ if T is odd. Combining the two inequalities above, we have

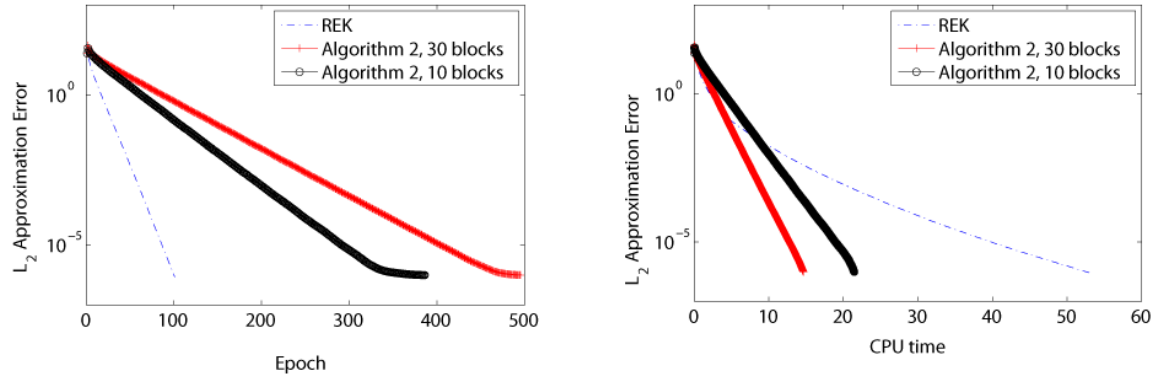


Figure 5: System with 1200×400 tomography matrix. Left: median ℓ_2 -norm error versus “epoch” (see footnote above). Right: median ℓ_2 -norm error versus CPU time.

$$\begin{aligned}
\mathbb{E} \|\mathbf{x}_T - \mathbf{x}_{\text{LS}}\|_2^2 &= \mathbb{E} \|\mathbf{x}_{k+k^*} - \mathbf{x}_{\text{LS}}\|_2^2 \\
&\leq \gamma^k \mathbb{E} \|\mathbf{x}_{k^*} - \mathbf{x}_{\text{LS}}\|_2^2 + \bar{\gamma}^{k^*} \frac{B}{1-\gamma} \\
&\leq \gamma^k \left(\gamma^{k^*} \|\mathbf{x}_0 - \mathbf{x}_{\text{LS}}\|_2^2 + \frac{B}{1-\gamma} \right) + \bar{\gamma}^{k^*} \frac{B}{1-\gamma} \\
&= \gamma^{k+k^*} \|\mathbf{x}_0 - \mathbf{x}_{\text{LS}}\|_2^2 + \left(\gamma^k + \bar{\gamma}^{k^*} \right) \frac{B}{1-\gamma} \\
&\leq \gamma^T \|\mathbf{x}_0 - \mathbf{x}_{\text{LS}}\|_2^2 + \left(\gamma^{\lfloor T/2 \rfloor} + \bar{\gamma}^{\lfloor T/2 \rfloor} \right) \frac{B}{1-\gamma}.
\end{aligned}$$

This completes the proof. □

References

- [Bjö96] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.
- [BN] J. Briskman and D. Needell. Block kaczmarz method with inequalities. *J. Math. Imaging Vis.* To appear.
- [BT87] J. Bourgain and L. Tzafriri. Invertibility of “large” submatrices with applications to the geometry of Banach spaces and harmonic analysis. *Israel J. Math.*, 57(2):137–224, 1987.
- [BT91] J. Bourgain and L. Tzafriri. On a problem of Kadison and Singer. *J. Reine Angew. Math.*, 420:1–43, 1991.

- [Byr08] C. L. Byrne. *Applied iterative methods*. A K Peters Ltd., Wellesley, MA, 2008.
- [CD12] S. Chrétien and S. Darses. Invertibility of random submatrices via tail decoupling and a matrix Chernoff inequality. *Statist. Probab. Lett.*, 82(7):1479–1487, 2012.
- [CEG83] Y. Censor, P. P. B. Eggermont, and D. Gordon. Strong underrelaxation in kaczmarz’s method for inconsistent systems. *Numer. Math.*, 41(1):83–92, 1983.
- [CFM⁺92] C. Cenker, H. G. Feichtinger, M. Mayer, H. Steier, and T. Strohmer. New variants of the POCS method using affine subspaces of finite codimension, with applications to irregular sampling. In *Proc. SPIE: Visual Communications and Image Processing*, pages 299–310, 1992.
- [CP12] X. Chen and A. Powell. Almost sure convergence of the Kaczmarz algorithm with random measurements. *J. Fourier Anal. Appl.*, pages 1–20, 2012. 10.1007/s00041-012-9237-2.
- [Deu85] F. Deutsch. Rate of convergence of the method of alternating projections. *Parametric optimization and approximation*, 76:96–107, 1985.
- [DH97] F. Deutsch and H. Hundal. The rate of convergence for the method of alternating projections, ii. *J. Math. Anal. Appl.*, 205(2):381–405, 1997.
- [EHL81] P. P. B. Eggermont, G. T. Herman, and A. Lent. Iterative algorithms for large partitioned linear systems, with applications to image reconstruction. *Linear Algebra Appl.*, 40:37–67, 1981.
- [Elf80] T. Elfving. Block-iterative methods for consistent and inconsistent linear equations. *Numer. Math.*, 35(1):1–12, 1980.
- [EN11] Y. C. Eldar and D. Needell. Acceleration of randomized Kaczmarz method via the Johnson-Lindenstrauss lemma. *Numer. Algorithms*, 58(2):163–177, 2011.
- [Gal05] A. Galántai. On the rate of convergence of the alternating projection method in finite dimensional spaces. *J. Math. Anal. Appl.*, 310(1):30–44, 2005.
- [GBH70] R. Gordon, R. Bender, and G. T. Herman. Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *J. Theoret. Biol.*, 29:471–481, 1970.
- [Han07] P. C. Hansen. Regularization tools version 4.0 for matlab 7.3. *Numer. Algorithms*, 46(2):189–194, 2007.
- [Her09] G. T. Herman. *Fundamentals of computerized tomography: image reconstruction from projections*. Springer, 2009.

- [HM93] G. Herman and L. Meyer. Algebraic reconstruction techniques can be made computationally efficient. *IEEE Trans. Medical Imaging*, 12(3):600–609, 1993.
- [HN90] M. Hanke and W. Niethammer. On the acceleration of kaczmarz’s method for inconsistent linear systems. *Linear Algebra Appl.*, 130:83–98, 1990.
- [HS78] C. Hamaker and D. C. Solmon. The angles between the null spaces of X-rays. *J. Math. Anal. Appl.*, 62(1):1–23, 1978.
- [Kac37] S. Kaczmarz. Angenäherte auflösung von systemen linearer gleichungen. *Bull. Internat. Acad. Polon.Sci. Lettres A*, pages 335–357, 1937.
- [KT94] B. Kashin and L. Tzafriri. Some remarks on coordinate restriction of operators to coordinate subspaces. Insitute of Mathematics Preprint 12, Hebrew University, Jerusalem, 1993–1994.
- [LL10] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: convergence rates and conditioning. *Math. Oper. Res.*, 35(3):641–654, 2010.
- [LWS14] J. Liu, S. J. Wright, and S. Sridhar. An asynchronous parallel randomized kaczmarz algorithm. *arXiv preprint arXiv:1401.4780*, 2014.
- [Nao11] A. Naor. Sparse quadratic forms and their geometric applications. Technical Report No. 1033, Séminaire Bourbaki, Jan. 2011.
- [Nat01] F. Natterer. *The mathematics of computerized tomography*, volume 32 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001. Reprint of the 1986 original.
- [Nee10] D. Needell. Randomized Kaczmarz solver for noisy linear systems. *BIT*, 50(2):395–403, 2010.
- [NT13] D. Needell and J. A. Tropp. Paved with good intentions: Analysis of a randomized block kaczmarz method. *Linear Algebra Appl.*, pages 199–221, 2013.
- [NW13] D. Needell and R. Ward. Two-subspace projection method for coherent overdetermined linear systems. *J. Fourier Anal. Appl.*, 19(2):256–269, 2013.
- [Pop98] C. Popa. Extensions of block-projections methods with relaxation parameters to inconsistent and rank-deficient least-squares problems. *BIT*, 38(1):151–176, 1998.
- [Pop99] C. Popa. Block-projections algorithms with blocks containing mutually orthogonal rows and columns. *BIT*, 39(2):323–338, 1999.
- [Pop01] C. Popa. A fast Kaczmarz-Kovarik algorithm for consistent least-squares problems. *Korean J. Comput. Appl. Math.*, 8(1):9–26, 2001.

- [Pop04] C. Popa. A Kaczmarz-Kovarik algorithm for symmetric ill-conditioned matrices. *An. Ştiinţ. Univ. Ovidius Constanţa Ser. Mat.*, 12(2):135–146, 2004.
- [PPKR12] C. Popa, T. Preclik, H. Köstler, and U. Rüdè. On Kaczmarz’s projection iteration as a direct solver for linear least squares problems. *Linear Algebra Appl.*, 436(2):389–404, 2012.
- [RT11] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. Available at [arXiv:1107.2848](https://arxiv.org/abs/1107.2848), Apr. 2011.
- [Sri10] N. Srivastava. *Spectral sparsification and restricted invertibility*. Phd dissertation, Yale University, New Haven, CT, 2010.
- [SS87] K. M. Sezan and H. Stark. Applications of convex projection theory to image recovery in tomography and related areas. In H. Stark, editor, *Image Recovery: Theory and application*, pages 415–462. Acad. Press, 1987.
- [SS12] D. A. Spielman and N. Srivastava. An elementary proof of the restricted invertibility theorem. *Israel J. Math.*, 190:83–91, 2012.
- [SV06] T. Strohmer and R. Vershynin. A randomized solver for linear systems with exponential convergence. In *RANDOM 2006 (10th International Workshop on Randomization and Computation)*, number 4110 in Lecture Notes in Computer Science, pages 499–507. Springer, 2006.
- [SV09] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15(2):262–278, 2009.
- [Tan71] K. Tanabe. Projection method for solving a singular system of linear equations and its applications. *Numer. Math.*, 17(3):203–214, 1971.
- [Tro08a] J. A. Tropp. Norms of random submatrices and sparse approximation. *C. R. Math. Acad. Sci. Paris*, 346(23-24):1271–1274, 2008.
- [Tro08b] J. A. Tropp. The random paving property for uniformly bounded matrices. *Studia Math.*, 185(1):67–82, 2008.
- [Tro09a] J. A. Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 978–986, Philadelphia, PA, 2009. SIAM.
- [Tro09b] J. Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 978–986. Society for Industrial and Applied Mathematics, 2009.

- [Ver01] R. Vershynin. John's decompositions: selecting a large part. *Israel J. Math.*, 122:253–277, 2001.
- [Ver06] R. Vershynin. Random sets of isomorphism of linear operators on Hilbert space. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 148–154. Inst. Math. Statist., Beachwood, OH, 2006.
- [WM67] T. M. Whitney and R. K. Meany. Two algorithms related to the method of steepest descent. *SIAM J. Numer. Anal.*, 4(1):109–118, 1967.
- [XZ02] J. Xu and L. Zikatanov. The method of alternating projections and the method of subspace corrections in Hilbert space. *J. Amer. Math. Soc.*, 15(3):573–597, 2002.
- [You12a] P. Youssef. A note on column subset selection. Available at `arXiv:1212.0976`, Dec. 2012.
- [You12b] P. Youssef. Restricted invertibility and the Banach–Mazur distance to the cube. Available at `arXiv:1206.0654`, June 2012.
- [ZF12] A. Zouzias and N. M. Freris. Randomized extended Kaczmarz for solving least-squares. *SIAM J. Matrix Anal. A.*, 34(2):773–793, 2012.